Print

Lesson 12: Confidence Intervals and Coefficient - Study Notes

Slide 1:

Confidence Intervals for Large Samples (n ≥ 30)

Sample sizes that are **smaller than 30** may require a slightly altered technique since the distribution cannot be assumed to be normal. We will look at small sample sizes later on in this lesson.

The quality of the estimation procedure is significantly altered by the sample statistic's variability and bias. A sample mean's variability is a function of the population's standard deviation and the sample size; it is the **standard error** of its sampling distribution.

$$SE = {}^{\sigma}/_{\sqrt{n}}$$

Therefore, the variability is minimized not only by a small standard deviation, but also by a large sample size (n). If the population's standard deviation is unavailable, the best estimate is the sample's standard deviation (s).

$$SE = {}^{s}/_{\sqrt{n}}$$

Slide 2:

Confidence Intervals for Large Samples (n ≥ 30) - Cont'd

When we use a sample mean to estimate the mean of a population, we know that although we are using a method of estimation, which has certain desirable properties, the chances are slim, virtually nonexistent, that the estimate will actually equal μ. Hence, it would seem logical to accompany such a point estimate of μ with some statement as to how close we expect the estimate to be. The error, $\bar{x}$ - μ, is the difference between the estimate and the parameter value. This difference is often referred to as a bias.

The relationship between the two characteristics can be described as follows:

$$Z^* = \frac{\bar{x} - \mu}{{}^{\sigma}/_{\sqrt{n}}} \quad \text{or} \quad Z^* = \frac{\bar{x} - \mu}{{}^{s}/_{\sqrt{n}}}$$

where "Z" is the converted value of a random variable ($\bar{x}$), which is approximately normally distributed. You may notice that this is very similar to the calculation of the z-score:

$$Z = \frac{x - \mu}{\sigma}$$

There are, however, two major differences:

- The z-score uses the difference between a particular random variable and the mean of the data set it came from (X - μ) or (X - $\bar{x}$), whereas the new equation deals with the difference between the sample mean and the population mean ($\bar{x}$ - μ).
- The standard error for the z-score calculation is simply the standard deviation. That is, the sample size (n) is always 1 because we are only looking at one score. Therefore, substituting 1 in the standard error equation yields:
  s / sqrt 1 = s / 1 = s.
  This would not be the case for sample sizes exceeding 1.

Slide 3:

Constructing the Interval

Now, taking a look at the new z-score formula, you may notice that all the variables can be determined by sample data, except one. The population mean, the one value that we want to estimate, is a variable in the formula!

Remember that z represents the number of standard deviations a random variable is from the population mean. If we rearrange the formula to solve for μ (keeping in mind that we want the sample mean to occupy the center of the interval), we would find that:

$$\mu = \bar{x} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

This formula will estimate the population mean for a given distribution (approximating the normal distribution) using a sample mean ($\bar{x}$), standard deviation (σ or s), and sample size (n), with 95% confidence.

Note: The actual value for z (for 95% confidence) is 1.96. We will see why shortly.

The **confidence level** (in this case: 95%) refers to the probability that the population parameter we are trying to find is actually contained within the interval. It is expressed as 1 - α, where α represents the allowable error, also known as the **significance level** (in this case: 5%).

Slide 4:

Constructing the Interval (Cont'd)

Example 1

You want to determine the amount of air in pounds per square inch (psi) that is pumped into a given tire by an automatic inflating machine. By repeating your experiment 40 times, you collect the following data:

- Amount of trials (n) = 40

- Sample Mean ($\bar{x}$) = 31.6 psi
- Sample Standard Deviation (s) = 2.2

Please construct a 95% confidence interval for the mean amount of air pumped into the tire by the automatic filler.

Answer:

The upper limit of the interval (UCL) = $\bar{x}$ + Z (s / sqrt n)
(N.B.: sqrt represents the "square root" so "sqrt 81" = 9)

31.6 + 1.96 (2.2/sqrt 40)
31.6 + 0.682
32.282

The lower limit of the interval (LCL) = $\bar{x}$ - Z (s / sqrt n)
31.6 - 1.96 (2.2/sqrt 40)
31.6 - 0.682
30.918

Therefore, we are 95% confident that the actual mean of the filling machine lies somewhere between **32.282** and **30.918** psi. The difference between the upper and the lower limit is known as the confidence interval's width.

Slide 5:

Constructing the Interval (Cont'd)

Example 2

A random sample of size n = 100 is taken from a population with σ = 5.1. Given that the sample mean is 21.6, construct a 95% confidence interval for the population mean μ.

Show Answer

Answer:

Substituting the given values of n, $\bar{x}$, s, and Z (for 95%) = 1.96 into the confidence-interval formula, we get:

$$\bar{x} - Z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

21.6 - 1.96 (5.1 / sqrt 100) < μ < 21.6 + 1.96 (5.1 / sqrt 100)
or
20.6 < μ <22.6

The interval from **20.6** to **22.6** contains the population mean μ, or it does not, but we are 95% confident that it does.

Slide 6:

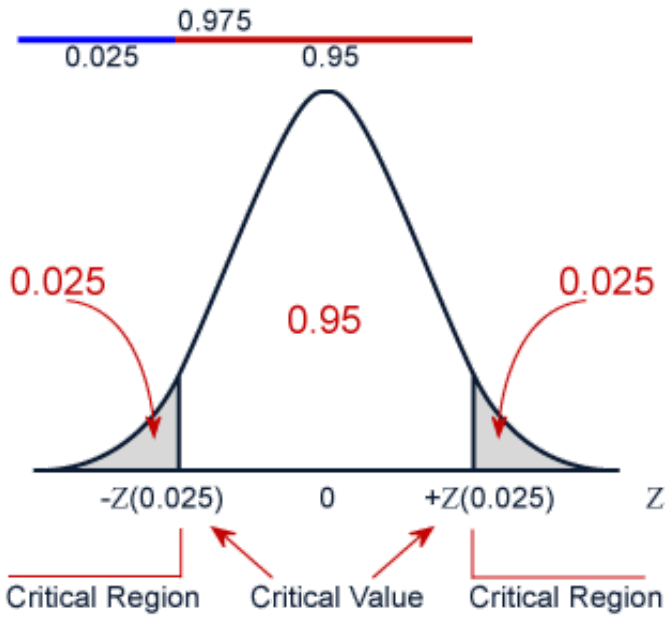
Determining Z

In the previous section, it was mentioned that the actual value for Z in a 95% confidence interval is 1.96. Now is the time to figure out how that was found.

First of all, let us look at the level of confidence as a confidence interval within the normal distribution. If the desired confidence level is 95%, then that area would be represented by the shaded region of the following graph:

The leftover 5%, known as the significance level (α), must be divided into two parts. These areas, found at the tails of the graph, represent the probability that the desired population parameter is found outside the interval's range. This extreme region is commonly called the **critical area**.

It is the situation represented here that must be reflected in the confidence interval formula as a function of "Z". We want to find the boundaries for our interval. If the significance level is known (or the confidence level), and the sample size is large (n = 30 or more), then we can determine the boundary with the aid of the normal distribution table.



Using the normal distribution, we want to find the Z value where 0.025 of the area is found (1) to the left, and (2) to the right.

This means that instead of starting with a Z-value and finding the corresponding area (which is the normal procedure), we will be doing the opposite. That is, given a particular area under the curve (i.e. 0.025), we must find the value within the table itself, and work our way backwards to the corresponding Z-value.

In this case, if we were to look up 0.025 on our normal distribution table (in the body of the table, not the borders), we would find that the corresponding Z-value is -1.96.

Since the regions are symmetric, we can also determine that the upper limit is 1.96. This value can be found by looking up the area 0.975, which came from adding the 0.95 to the 0.025, which is all the area to the left of $Z_{(0.025)}$

The value of ± 1.96 represents the critical value of the confidence interval and is expressed as a function of Z in the following notation:

$$\pm Z_{(\alpha/2)}$$

This is known as the **confidence coefficient** or **critical value**. It is interpreted as the Z-value with an area of α/2 to the left of its negative value - $Z_{(\alpha/2)}$ on the normal distribution, and α/2 to the right of $Z_{(\alpha/2)}$. It serves as the boundary for the critical region (the region that represents Ha).

Slide 7:

Determining Z (Cont'd)

In summary, the final formula for determining a confidence interval is:

$$\bar{x} - Z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} : \text{Lower confidence limit (LCL)}$$

$$\bar{x} + Z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} : \text{Upper confidence limit (UCL)}$$

$$\bar{x} - Z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

Note that σ can be replaced by S if unknown.

Using the same rationale, can you confirm that the $Z_{(\alpha/2)}$ values for other popular confidence levels are the following?

Confidence Level	$Z_{(\alpha/2)}$
99%	**2.575**
98%	**2.33**
95%	**1.96**
90%	**1.645**

When using the normal distribution table, it is quite possible that an exact value we are seeking cannot be found. If that is the case, you have two options:

- Use the closest value possible (as in 98%, 2.33 gave 0.9901, which is closest to 0.9900).
- Use the average of the two closest values (as in 90%, used average between 1.64 (0.9495) and 1.65 (0.9505)).

Slide 8:

Error and Sample Size

Since the whole point of a confidence interval is to predict an unknown value, error is not only common, it is acceptable. The error that we allow in our estimates can, however, be controlled by manipulating certain variables that are directly associated with it.

The maximum error of an estimate, E, is the product of $Z_{(\alpha/2)} and \ {}^{\sigma}/_{\sqrt{n}}$.

$$E = Z_{(\alpha/2)}\left({}^{\sigma}/_{\sqrt{n}}\right)$$

The confidence interval's width (upper limit - lower limit) can also be expressed as a function of E:

CI width = 2E

Therefore, the width of the interval is directly dependent on the interval's confidence level (expressed by the confidence coefficient), the variability (expressed by "σ"), and on the sample size. Since the variability is not malleable, the researcher can manipulate the width of the CI by altering the confidence level and sample size.

Slide 9:

Error and Sample Size (Con't)

Example 1

We are interested in determining the mean height (in cm) of women between the ages of 25 and 29.

Random sample of n = 100 women between the ages of 25 and 29.

- sample mean = 165 cm
- sample standard deviation s = 5 cm
- standard error of the mean (SE) = 5 / (sqrt) 100 = 0.5

The 95% confidence limits for the population mean, μ are:

- 165 + 1.96 (0.5) = 165 + 0.98 = **165.98**
- 165 - 1.96 (0.5) = 165 - 0.98 = **164.02**

i.e. 95% confidence interval for μ is (164.02, 165.98) cm.

Slide 10:

Interpretation of the Confidence Interval

1. The sample mean provides a point estimate (i.e. single value approximation) for μ.
   - e.g. in the example about women's height = 165 cm, μ is approximately 165 cm.

2. Confidence limits provide an interval estimate together with a degree of confidence (**accuracy**) that the parameter is in the interval.

- e.g. with 95% confidence the population mean height μ for these women is in the interval (164.02, 165.98) cm.

3. The width of the interval (i.e. **precision** of estimate) depends on the sample size.

- e.g. in the example about women's height the sample size was n = 100 so the 95% interval is 165 ± 1.96 (0.5) = (164.02, 165.9).

Slide 11:

Interpretation of the Confidence Interval (Cont'd)

Suppose the sample size for the previous problem had been n = 40, but the mean and standard deviation were unchanged (165 and 5). The standard error would then be: SE (μ) = 5 / (sqrt) 40 = 0.791.

Then a 95% confidence interval for μ is 165 ± 1.96 (0.791) = 165 ± 1.55, which gives (163.5, 166.5).

Notice that increasing the sample size increases the precision of the estimate by decreasing the width of the interval.

e.g. width of 95% confidence interval = Upper Limit - Lower Limit

$$= \left(1.96 * \left(\sigma/\sqrt{n}\right)\right) + \left(1.96 * \left(\sigma/\sqrt{n}\right)\right)$$

$$= approx.\ 2 * 1.96 * \left(\sigma/\sqrt{n}\right)$$

So, if n = 100, width = 2 x 1.96 x (5/sqrt 100) = **1.96**
or if n = 25, width = 2 x 1.96 x (5/sqrt 25) = **3.92**

In this case, if you increased the sample size by a factor of 4 you would decrease the width of the confidence interval by 2.

The **precision** of the estimate depends on the standard error [SE = s / (sqrt n)], whereas the **accuracy** depends on the confidence coefficient ($Z_{\alpha/2}$).

If we wanted to determine the sample size needed in order to obtain a certain level of error, we could rearrange the formula for E and find:

$$n = \left[\frac{Z_{(\alpha/2)} * \sigma}{E}\right]^2$$

Slide 12:

Interpretation of the Confidence Interval (Cont'd)

Example 2

A statistics teacher intends to use the mean of a random sample of size n = 150 to estimate the average student's statistics knowledge (as measured by a certain test) of students enrolled in the Exercise Science program. If, on the basis of experience, the teacher can assume that s = 6.2 for such data, what can he assert with a probability of 0.99 about the maximum size of his error?

Show Answer

Answer:
$\alpha = 1 - 0.99 = 0.01$, $\alpha/2 = 0.005$
$Z\alpha/2 = Z_{(0.005)}$, so to find z, we must find 0.005 within the body of the normal distribution table. You will find 0.005 at the intersection of the row -2.5 and the columns 0.07 and 0.08. Since we are as close to one side as to the other, z = 2.575.

Substituting n = 150, s = 6.2, and $Z_{(0.005)} = 2.575$ into the formula for determining E, we get
E = 2.575 x (6.2/ sqrt 150) = 1.30.
Thus, the teacher can assert with a probability of 0.99 that his error will be at most 1.30.

Suppose now that the statistics teacher collects his data and gets a sample mean of 69.5. Can he still assert with a probability of 0.99 that the error is at most 1.30? After all, the sample mean differs from the true value at most 1.30.

Well, he can, but it must be understood that the 0.99 probability applies to the method he used to determine the maximum error (getting the sample data and using the formula for E) and not directly to the parameter he is trying to estimate. To make this distinction, it has become customary to use the word "confidence" here instead of "probability." In general, we make probability statements about the future value of random variables (say, the potential error of an estimate) and confidence statements once the data have been obtained. Accordingly, we would say in our example that the statistics teacher can be 99% confident that the error of his estimate, $\bar{x} = 69.5$, is at most 1.30.

Slide 13:

Interpretation of the Confidence Interval (Cont'd)

Example 3

Please determine the sample size needed to estimate the mean height (in inches) of all first-year male university students in Quebec if we want a maximum error of 1 inch with 95% confidence. Assume that the population is normally distributed and that its standard deviation is 3 inches.

Show Answer

Answer:
Using the formula for sample size, substitute the $Z_{(\alpha/2)}$, standard deviation, and E with 1.96, 3, and 1. Squaring the result yields 34.57. Since the sample size is a discrete variable (we cannot sample 0.57 of an individual!) we round the number up to a whole number. Therefore, we would need a sample size of 35 to satisfy our criteria.

Slide 14:

The T-Distribution

Inferences about a population parameter are based on the statistics from a sample data set whose sampling distribution of means is normally distributed. If the shape of the distribution is unknown, it can be assumed that it is normally distributed if the sample size is large (n >= 30). But what happens if we are faced with a situation where the sample size is small (n < 30) and the shape of the distribution is unknown?

If this is the case (n < 30), then our next point of interest lies with the population standard deviation (σ). If the parameter is known, **which is very rare**, we can assume a normal distribution and continue with z-score calculations and the standard normal distribution table. When σ is unknown, but n is large, we are able to estimate σ using the sample standard deviation (s), and we have no problem using the normal distribution.

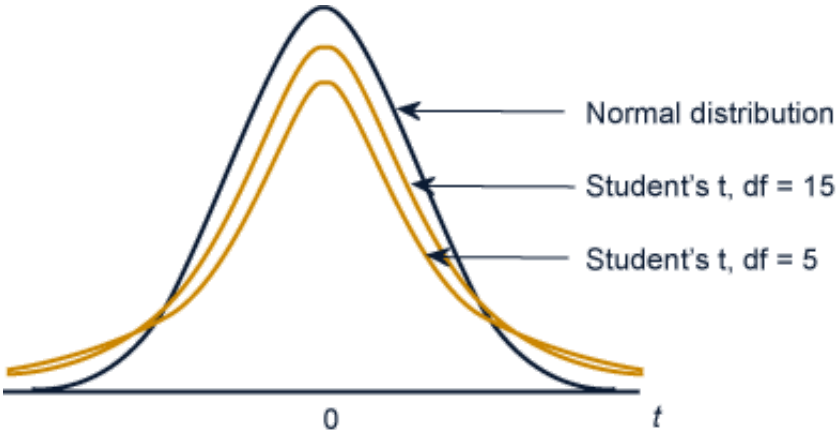
However, should (1) the sample size be small (n < 30) AND (2) the population standard deviation (σ) is unknown, then we cannot assume a standard normal distribution. Instead, we must employ a similar distribution known as the t-distribution.

Slide 15:

The T-Distribution (Cont'd)

Properties of the T-distribution

- It is symmetrically distributed about its mean of 0.
- It has a variance greater than 1, but as n increases, the variance approaches 1.
- Each sample size forms its own distribution such that as n increases, the shape of the distribution narrows and increases in height until n = 30, in which case, it is considered approximately normally distributed.
- To compensate for underestimations, the t-distribution uses degrees of freedom (n - 1).



Notice the physical differences between each sketch due to the sample size. Each degree of freedom causes a different distribution. The smaller the sample, the wider and less-peaked the distribution (since there is less variance in the values).

Note: The t-distribution is sometimes called Student's t-distribution. "Student" is the pseudonym used by William Gosset when he published his work in 1908 while working at the Guinness

Brewery in Dublin, Ireland. Apparently he was not allowed to publish scientific papers while working for the company, hence the need to conceal his identity.

Slide 16:

Reading the T-Distribution Table

Upon inspecting the t-distribution table, you will notice that it is a little different than the standard normal distribution table. First of all, the left-hand side of the table starts at 1 and increases systematically until 29, and the last value is infinity (inf). This column represents the **degrees of freedom (n-1)**. The next columns represent the **significance level** (α). You will notice that we are limited to 5 values (0.1, 0.05, 0.025, 0.01, and 0.005). Furthermore, the values found throughout the table are all positive and greater than 1.

The value found at the intersection of the degrees of freedom and the significance level represents the **critical value** for the t-distribution based on that data. It is this value that will serve as the boundary for the critical region of the distribution, or as the **confidence coefficient**, as did the $Z_{(\alpha/2)}$ in confidence intervals. In order to use the value from the t-table in a confidence interval, we must consider the fact that significance level is halved, just as it was when we used the normal distribution. Therefore, this must be reflected when we look for our value in the t-table.

Let us say, for argument's sake, that we are dealing with a confidence interval from a sample size of (n) 25 and a significance level of (α) 0.05. Assuming that the population standard deviation is unavailable, we must turn to the t-table. Our first step is to determine the degrees of freedom, which is simply the sample size minus one (n - 1) = 24. Then we must identify the significance level (0.05) and divide it in 2, just as we did with $Z_{(\alpha/2)}$. That means that instead of looking up an α of 0.05, we use 0.025. Therefore, our confidence coefficient is 2.064.

The proper notation for this is:

$$t_{(n-1,\alpha/2)}$$

Slide 17:

Confidence Interval for Small Sample Sizes (σ unknown)

The procedure for confidence intervals when the sample size is small (< 30) and the population standard deviation is unknown varies only in the determination of the confidence coefficient. Instead of using the normal distribution to find Z (α/2), we use the t-distribution to find t (α/2) and we simply substitute the value into a modified equation for the confidence interval.

$$\bar{x} - t_{(n-1,\alpha/2)}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(n-1,\alpha/2)}\frac{s}{\sqrt{n}}$$

Slide 18:

Confidence Interval for Small Sample Sizes (σ unknown) (Cont'd)

Example 3

A sample of n = 10 bottles of a leading brand of soft drink were measured and the contents were (in ml):

299	276	283	301	297
281	300	291	295	291

Determine the mean amount of liquid found in a typical bottle with 90% confidence.

Answer:

From the sample data, we found that:

* Sample mean = 291.4 ml
* Sample standard deviation = 8.72 ml
* Sample size = 10
* Confidence coefficient for 90% = $t_{(df,\ \alpha/2)} = t_{(10-1,\ 0.1/2)} = t_{(9,\ 0.05)} = 1.833$.

Substituting the values into the new equation yields:

Upper Limit = 291.4 + 1.833 * 8.72 / sqrt 10 = 296.45
Lower Limit = 291.4 - 1.833 * 8.72 / sqrt 10 = 286.35

Therefore, 286.35 < μ < 296.45.

Slide 19:

Recap

Depending on the sample size and desired confidence level, the confidence coefficient is a factor that is pulled off either from the normal distribution or from the t-tables. This value is used in conjunction with the standard error to make an interval estimate of a parameter.

- The standard error is a ratio involving the standard deviation and the sample size (n).
- The confidence level of a confidence interval is directly related to the significance level such that 100% - α = confidence level.
- The confidence coefficient is pulled off the normal distribution table if the sample size is greater than or equal to 30. Otherwise, we use the t-table.
- There is a direct inverse relationship between sample size and error in the estimate. As the sample size increases, the error goes down.
- The minimum sample size can be estimated based on the desired significance level and accepted error.
- Regardless of the sample size, the general formula needed to construct a confidence

interval remains the same.

You can post a message online in your discussion folder any time you have something to share with your discussion group concerning the current lesson. Simply click Discussion Board or use the menu at the top of the screen.

Next lesson: Hypothesis Testing